Factors Influencing Student Success

Yenmin Young
Springboard Data Science
Capstone 3 Project

Project Overview

Problem Statement:

- What factors most influence academic success?
- Why do some students pass while others fail? Can we predict who is at risk?

Context: Education outcomes shape future access to jobs, higher education, and social mobility. In Portugal, two datasets capture a range of student academic and lifestyle data.

Success Metrics: RMSE for grade prediction; Accuracy, F1, and ROC-AUC for pass/fail classification

Stakeholders

- **Students** : Jnsights can empower them to take action
- **Families** : Want to support their children
- **Educators** : Need early warning tools
- **Policymakers** : Can allocate resources more effectively

Constraints

- Observational data: Cannot establish causality
- Limited to one country's education system
- Potential bias in self-reported data (e.g., alcohol use)

Main Goals

- Predict whether a student will pass/fail
- Identify key features that influence final grade

Approach

- Classification for pass/fail outcome (binary)
- Regression for final grade prediction (continuous)

Dataset Overview

Data Source: Two datasets from Portuguese secondary schools (Math & Portuguese classes)

Features Include:

- Demographics (age, gender, parental education)
- Academics (grades, study time, absences, failures)
- Lifestyle (alcohol use, work, relationships, extracurriculars)

Target Variables: Final grades (G3); Binary indicator for pass/fail (10+ to pass)

Data Wrangling

Steps Taken:

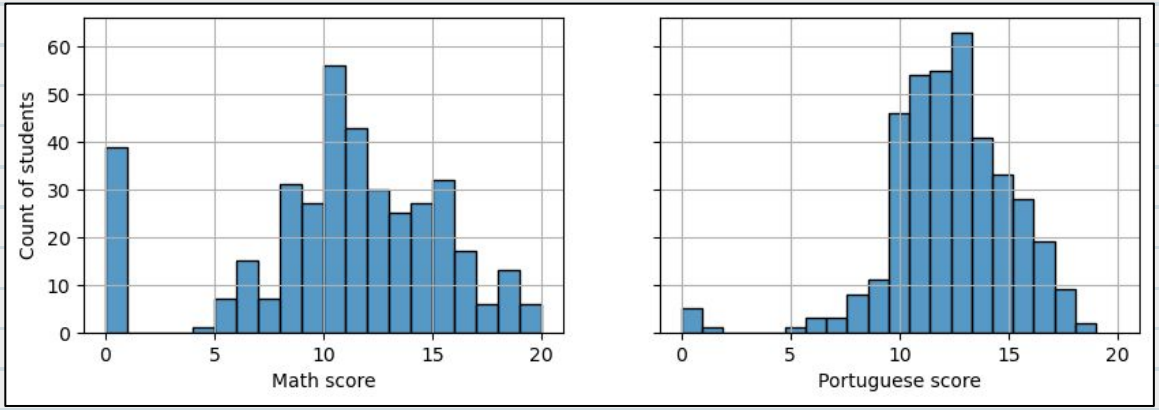
- Merged math and Portuguese datasets
- Verified no duplicates and no missing values
- One-hot encoded categorical variables
- Scaled data and performed PCA for dimensionality reduction
- Added pass/fail binary column

Key Insight: Clean dataset with rich, diverse features, ideal for supervised learning

Exploratory Data Analysis

Grade Distributions Findings:

- 246 students passed both classes
- 104 failed Math but passed Portuguese
- 23 failed both
- Math has significantly more failures



Questions Raised:

- Why is Math harder to pass?
- Why do some students drop to zero mid-term? (see next slide)

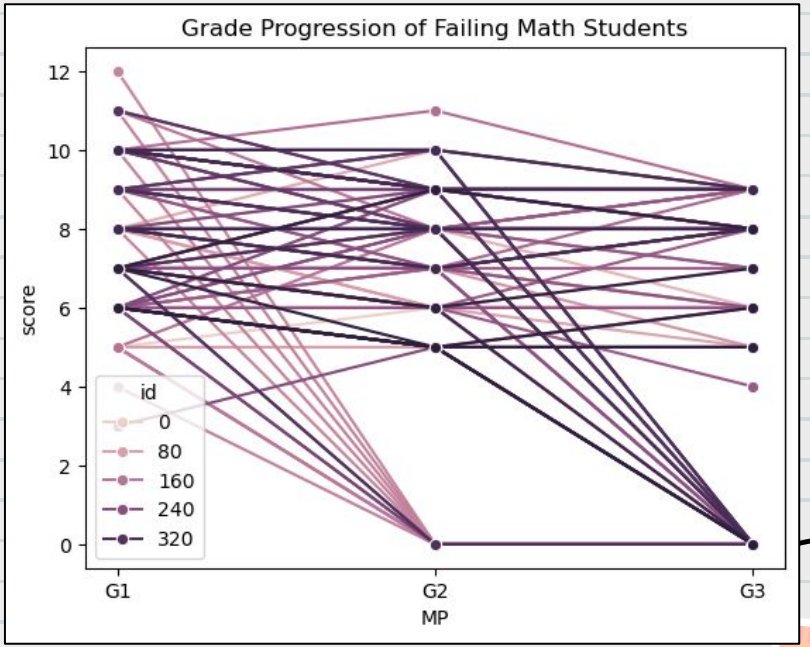
Exploratory Data Analysis

Student Behavior Patterns Patterns Identified:

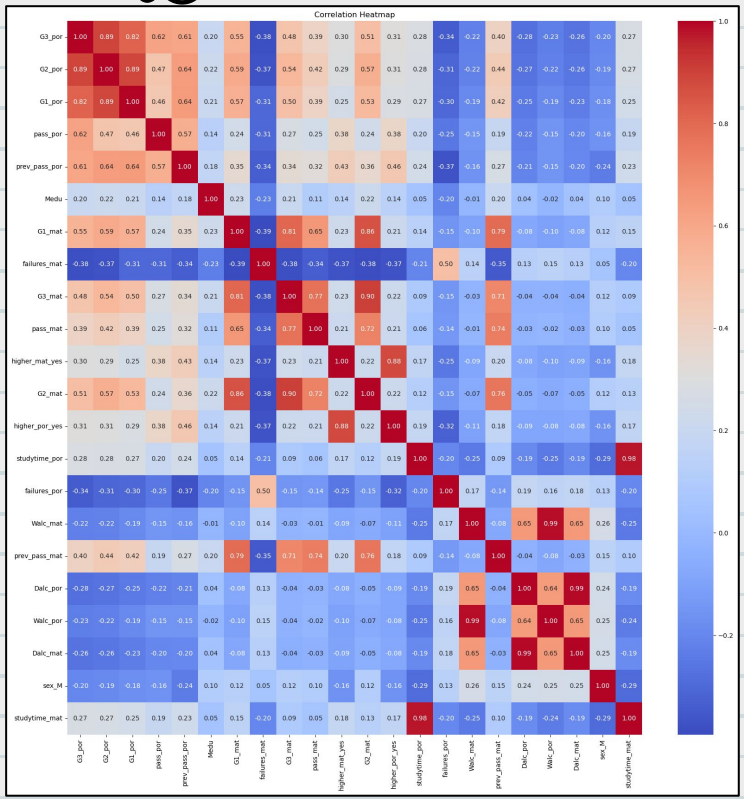
- Sudden drops to 0 in G2 or G3 for some students
- Consistently low performers vs. sudden decline

Next Steps:

- Investigate grade policies, life events, or systemic biases



Exploratory Data Analysis



Correlation Heatmap Highly Correlated Features:

- **Positive** : study time, higher education plans, mother's education level
- **Negative** : weekday/weekend alcohol use, number of past failures

Modeling Approach

Regression (Final Grade Prediction):

- Tested: Linear, Ridge, Lasso, RF, Gradient Boosting, XGBoost
- **Best: Gradient Boosting**
  - Math RMSE: 4.11
  - Portuguese RMSE: 4.22

Classification (Pass/Fail):

- Tested: Logistic, RF, Gradient Boosting, XGBoost, KNN, SVC
- **Best: Random Forest**
  - Math Accuracy: 0.745
  - Portuguese F1: 0.887, ROC-AUC: 0.948

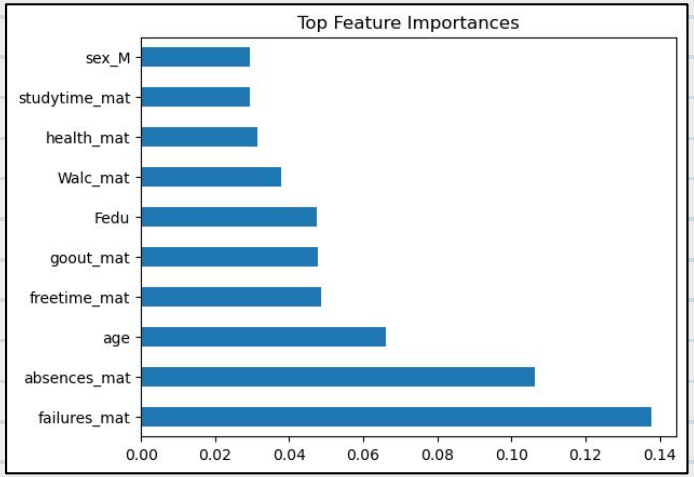
Why These Models?

- **Gradient Boosting:** Handles complex nonlinear relationships and interactions well, strong performance on small datasets
- **Random Forest:** Robust to overfitting, handles feature importance ranking well, high accuracy

Analysis Results

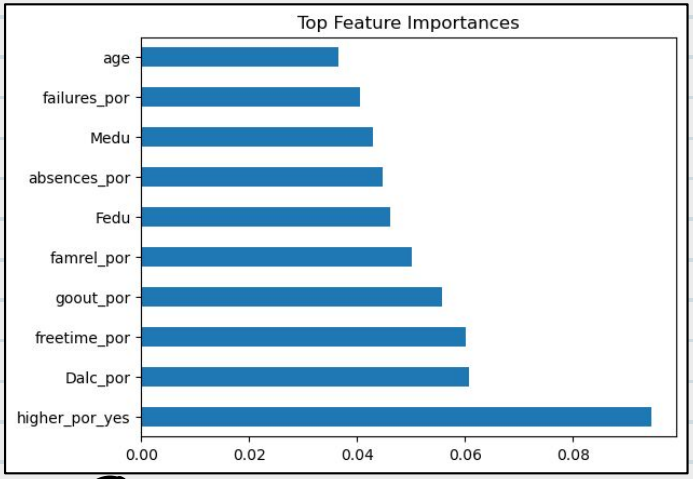
Top Features (Math):

- Past failures
- Number of absences
- Age, free time, going out



Top Features (Portuguese):

- Plans for higher education
- Weekday alcohol use
- Family relationship quality



Recommendations

- Flag students with high absences or past failures for early intervention
- Promote higher education awareness and mentorship programs
- Launch alcohol awareness campaigns in schools
- Provide family engagement opportunities
- Strengthen math-specific academic supports

Further Research

Investigate Deeper

- Interview teachers/students to explain sudden grade drops
- Identify biased or unfair grading and absence policies

Improve Research

- Expand dataset to include additional schools/regions
- Integrate mental health, work obligations, and home responsibilities into models
- Explore longitudinal modeling over multiple years